ARGUMENTA OECONOMICA
No 1 (46) 2021
PL ISSN 1233-5835

I. ARTICLES

*Marta Małecka**

INDUSTRY STANDARD AND ECONOMETRIC STANDARD: THE SEARCH FOR POWERFUL APPROACH TO EVALUATE *VaR* MODELS

Under the Basel III and Basel IV accords, risk model validation remains based on the *VaR* measure. According to the industry practice, *VaR* backtesting procedures rely on two likelihood ratio tests, which, in light of the academic research, have been criticized for their unsatisfactory power. This paper aims to show the differences between *VaR* model evaluation based on the standard likelihood ratio approach and backtesting by means of other econometric methods applicable to the binary *VaR* failure process. The author decomposed the model evaluation into testing the unconditional coverage, replaced the likelihood ratio with a normal statistic, and in the next stage in order to verify the conditional coverage, employed the Ljung-Box statistic. The study experimentally confirmed the superiority of the proposed procedures over the industry standards. The main contribution, however, is the empirical study designed to demonstrate the practical differences in risk analysis attributable to the choice of the backtesting method. Using data on leading stock market indexes, from various periods, the author showed that the practical conclusions from backtesting diverge markedly due to the test choice. The proposed, more powerful tests, contrary to the standard procedures, allowed for distinguishing distinct models of index behaviour connected with undergoing the financial crises.

Keywords: *Value-at-Risk*, backtesting *VaR*, Kupiec test, Markov test, test size, test power
JEL Classification: C22, C52, D53, G11
DOI: 10.15611/aoe.2021.1.01

1. INTRODUCTION

Value-at-Risk (*VaR*) owes its popularity as a risk measure to both business practice and international supervisory rules. In the context of business routines, its constantly widening range of applications stems from the practical advantages, like the straightforward interpretation and applicability to complex portfolios. Following the industry practice, the international system of risk measurement standards was based on *VaR* in the 1990s (Basel 1996), shortly after the original inception of this measure by J. P. Morgan (1994). Although

* Department of Statistical Methods, University of Lodz, Poland.

the reform of the supervisory rules, undertaken in 2012 by the Basel Committee of Banking Supervision (Basel 2016, 2017) has involved a movement from *VaR* to the *ES* (Expected Shortfall) measure, the procedures of risk model evaluation remain based on *VaR*. This necessitates a discussion on *VaR* testing rules and gives an incentive to investigate the statistical properties of relevant methods.

The *VaR* testing framework is based on a binary variable indicating *VaR* violations. Under the correct risk model this variable is required to follow the iid Bernoulli process. The iid Bernoulli property is commonly split into the postulates of the unconditional and conditional coverage property. The first postulate refers to the overall *VaR* failure rate and means that the number of violations should match the assumed *VaR* tolerance level, while the conditional coverage property requires the independence of violations. The extensive toolkit for verifying these two postulates, separately or jointly, involves testing the parameters of the Bernoulli process (Kupiec 1995), using the transition probabilities of the binary Markov chain (Christoffersen 1998), regressing *VaR* failures on their lagged values (Engle and Manganelli 2004), checking the unpredictability of the durations between *VaR* failures (Christoffersen, Pelletier 2004, Candelon et al. 2011) or using the spectral theory (Berkowitz et al. 2011, Gordy and McNeil 2018). The one-level *VaR* backtesting procedures were extended into checking the fit of the density function (Berkowitz 2001), the truncated density function (Crnkovic and Drachman 1997) or multi-level *VaR* testing (Hurlin and Tokpavi 2007, Colletaz et al. 2013, Kratz et al. 2018). Among these propositions, two tests, formulated within the likelihood ratio (LR) framework, have won wide recognition in the industry. These are the Kupiec test (Kupiec 1995), which checks the unconditional failure rate, and Christoffersen's Markov test (Christoffersen 1998), aimed at capturing the serial dependence in failures. Developed specifically for the purposes of risk management, these tests offer the advantage of a convenient, straightforward implementation to real-life processes. These popular approaches, however, have been repeatedly criticized with respect to their statistical properties (Lopez 1999, Christoffersen and Pelletier 2004, Berkowitz et al. 2011, Pajhede 2017).

In view of practical aspects, like the straightforward implementation and computational efficiency, the study explored the possibilities of backtesting *VaR* through standard econometric methods, applicable to the Bernoulli sequence and independence testing. Building on the results of Malecka (2018), the author refrained from using methods developed specifically for the purposes of risk management. Exploiting the properties of the Bernoulli distribution, the binomial distribution and its convergence to the normal one,

the study applied a normal statistic to checking the unconditional coverage property. To enhance the power properties in the conditional coverage testing, the author employed the Ljung-Box statistic (Ljung and Box 1978). This used the fact that the Ljung-Box test has the power against linear alternatives of any order, which corresponds to dependencies in the *GARCH* processes, with a slow decay of correlation.

The aim of the paper was to evaluate the capacity of the above-mentioned, well-established econometric methods as risk management tools, in relation to the popular *VaR*-dedicated tests. For this purpose the Monte Carlo technique was employed, and an empirical investigation of the methods was performed. The Monte Carlo study was designed to reflect the typical *VaR* failure setting. To achieve this, the study used two types of experiments. In the first type, correlated *VaR* violations were generated by employing the *GARCH*-class models with the specification that enables explicit indication of the volatility clustering. Therefore it was possible to study the power of the tests as a function of a controlled parameter of a return distribution. However, the explicit control over the parameter that represents volatility clustering, limits the range of applicable data generating processes. Therefore this type of experiment was followed by the second type, where the priority was to closely reflect the real-life financial processes. The second set of experiments stepped away from the exact control over the volatility clustering and, instead, used the *ARMA* and *N-GARCH*-based data generating processes with the Student *t*-distribution and parameters based on empirical data. In accordance with the Basel framework, the study provided the results for *VaR* coverage levels 1% and 2.5%, in this way extending the earlier research on backtesting *VaR* through classical econometric methods, which treated only 5% *VaR* (Malecka 2018). The author experimentally showed that the proposed approach outperforms the standard LR tests both in terms of the accuracy, understood as a test size, and in terms of ability to detect incorrect risk models, understood as a test power. The main contribution is, however, abroad empirical study, which exploits and illustrates the results of the Monte Carlo simulations. The research was designed to show the differences in risk analysis that result from the choice of a backtesting procedure. To this end, three leading stock market indexes were utilised, for which the standard LR and the proposed methods were subsequently applied. To provide a relevant scenario for assessing the capacity of risk management tools, all the backtesting procedures were implemented to evaluate twelve mainstream market risk models in various periods, under diverse volatility conditions. The results demonstrate that the proposed tests, compared to the popular LR approach, provide a more insightful view of market behaviour. They allow to choose models suitable for

predicting risk under various volatility regimes and thus characterize the market specificity. Contrary to the standard procedures, they distinguish two distinct patterns of the market behaviour, connected with undergoing the financial crises.

The paper proceeds as follows. Section 2 sets the notation and provides the details of the compared tests. Section 3 gives the comparative evaluation of their properties through the Monte Carlo experiments, based on the *GARCH* processes. Section 4 empirically illustrates the differences in the outcomes of real data analysis, resulting from the choice of testing procedure.

2. TESTING UNCONDITIONAL AND CONDITIONAL *VaR* COVERAGE

The *VaR* model evaluation framework is based on a binary process indicating *VaR* failures. Assuming that R_t is the random return from a portfolio, with the continuous distribution function F_{R_t}, and *VaR* is its p-quantile, $VaR_p(R_t) = F_{R_t}^{-1}(p)$, the failure process is defined as:

$$I_t = 1_{\{R_t < VaR_p(R_t)\}}. \tag{1}$$

The quantile order p is referred to as the *VaR* tolerance level. Under the correct *VaR* model, the I_t process is required to be the iid Bernoulli process with the parameter p, i.e. $I_t \overset{iid}{\sim} B(p)$. The iid Bernoulli condition may be decomposed into the postulate of unconditional coverage, referring to the unconditional probability of failure p, and the postulate of conditional coverage, requiring independence of failures.

The industry standard to test the unconditional coverage property is the Kupiec test (Kupiec 1995), which assumes the identical, independent Bernoulli distribution $I_t \overset{iid}{\sim} B(\pi_1)$ and checks the p_1 parameter: $H_0 : \pi_1 = p$. The parameter value is estimated through the empirical rate of violations $\hat{\pi}_1 = \frac{T_1}{T}$, where T_1 is the number of violations and T is the number of observations. The H_0 restriction is checked through the likelihood ratio statistic:

$$LR_{uc} = -2\log\frac{p^{T_1}(1-p)^{T_0}}{\hat{\pi}_1^{T_1}(1-\hat{\pi}_1)^{T_0}}, \tag{2}$$

Where $T_0 = T - T_1$. With one parameter restriction, the likelihood ratio LR_{UC}, under the null, has the asymptotic $\chi^2_{(1)}$ distribution.

Under the above assumptions, the T_1 statistic, as the sum of iid Bernoulli variables: $T_1 = \sum_{t=1}^{T} I_t$, has the binomial distribution $T_1 \sim B(T, \pi_1)$, which, provided that H_0 restriction is satisfied, changes into $T_1 \sim B(T, p)$, If the number of observations is large, by the Central Limit Theorem the binomial distribution converges to the normal one. Exploiting this fact, the unconditional coverage *VaR* test may also by conducted with the use of the continuous normal distribution. The test statistics *Z* takes the form:

$$Z = \frac{T_1 - Tp}{\sqrt{Tp(1-p)}} \tag{3}$$

and, under the null, has the asymptotic standard normal distribution $N(0,1)$.

The unconditional coverage tests, relying on the iid assumption, consider only the overall rate of failures. The complete *VaR* backtesting procedure, as formulated in the conditional coverage postulate, requires also checking independence of failures. The standard approach to verifying the conditional coverage property is the Markov test (Christoffersen 1998), which embeds the failure process within the binary first-order Markov chain. The test is formulated in terms of single-step transition probabilities. The independence condition implies that the transition probabilities π_{01} and π_{11} are equal, where π_{ij} denotes the probability of the transition of I_t from state *i* to state *j*. The null $H_0 : \pi_{01} = \pi_{11}$ is tested through the likelihood ratio statistic of the form:

$$LR_{cc} = -2\log \frac{\hat{\pi}_1^{t_1}(1-\hat{\pi}_1)^{t_0}}{\hat{\pi}_{01}^{t_{01}}(1-\hat{\pi}_{01})^{t_{00}}\,\hat{\pi}_{11}^{t_{11}}(1-\hat{\pi}_{11})^{t_{10}}}, \tag{4}$$

where $\hat{\pi}_1 = \frac{T_1}{T_0 + T_1}$, $\hat{\pi}_{01} = \frac{T_{01}}{T_0}$, $\hat{\pi}_{11} = \frac{T_{11}}{T_1}$ and T_{ij} is the empirical number of transitions from state *i* to state *j*. The likelihood ratio LR_{CC}, under the null, has the asymptotic $\chi^2_{(1)}$ distribution.

Relying on the single-step transition probabilities, the Markov test has only the potential to detect first-order dependencies. This deficiency may be made up for by employing the well-known econometric Ljung-Box test, which has the power against linear alternatives of any order. The application of the Ljung-Box test to a *VaR* failure process implies the null formulated in terms of correlation coefficients between *VaR* violations, $H_0 : \rho_h = 0$, $h = 1, 2, \ldots, H$, $H < T$. Then the test statistic has the following form:

$$LB_H = T(T+2)\sum_{h=1}^{H}\frac{\hat{\rho}_h^2}{T-h}, \tag{5}$$

where $\hat{\rho}_h$ are sample autocorrelations of order h in the I_t process. Under the null, the LB_H statistic has the χ^2_H distribution.

3. STATISTICAL PROPERTIES OF *VaR* BACKTESTING PROCEDURES

As a preview to the empirical analysis, a Monte Carlo study was used to assess the theoretical statistical properties of the examined tests in the context of the *VaR* model evaluation. The comparative study included the size and power properties, estimated as the proportion of rejections under the null and under the alternative, respectively. The size evaluation included significance levels 0.01, 0.05 and 0.1. For the power comparison, the study reported rejection frequencies at 0.05 level. The estimates of the statistical properties were computed over 10,000 Monte Carlo trials for sample sizes 100, 250, 500, 750, and 1000[1].

With reference to the unconditional coverage property, the author compared the normal Z statistics, applied to a *VaR* failure series, to testing *VaR* models through the Kupiec LR_{UC} test. In accordance to the conditional coverage property the author evaluated the properties of the Ljung-Box LB_H in relation to the properties of the Markov-chain-based LR_{CC} procedure, setting the autocorrelation order $H = 5$, which corresponds to one week of daily observations[2].

The size study investigates test accuracy, understood as the compliance between the observed rejection frequency and the nominal significance level (Tables 1 and 2). Since the size assessment examines the test performance under the null, it requires data from the iid binary process with the correct failure probability. This was done through generating iid Bernoulli samples with the parameter π_1, equal to the chosen *VaR* tolerance level.

While the size results in the group of the unconditional coverage tests show minor differences between the compared methods, the discrepancies observed between the conditional coverage tests are much larger. Both unconditional tests – Z and LR_{UC} – seem relatively accurate, however any differences that

[1] All computations in this study are conducted with the use of the MATLAB software.

[2] The power of the LB test with respect to the choice of the autocorrelation order was studied by Berkowitz et al. (2011) and Pajhede (2017). Via simulations, Berkowitz et al. (2011) showed that the autocorrelation order H=5 outperforms H=1. In similar simulations, Pajhede (2017) argued that H=5 outperforms H=10. Using both these results, the author chose H=5.

Table 1

Size estimates of unconditional coverage *VaR* tests

Test	Significance level	1% *VaR*					2.5% *VaR*				
		Sample size					Sample size				
		100	250	500	750	1000	100	250	500	750	1000
LR_{UC}	0.01	0.006	0.005	0.006	0.009	0.013	0.005	0.008	0.010	0.013	0.012
	0.05	0.027	0.015	0.067	0.040	0.056	0.014	0.056	0.058	0.057	0.058
	0.1	0.027	0.048	0.067	0.100	0.114	0.043	0.108	0.106	0.115	0.116
Z	0.01	0.017	0.014	0.006	0.016	0.010	0.012	0.009	0.011	0.010	0.010
	0.05	0.078	0.044	0.038	0.061	0.038	0.043	0.039	0.055	0.046	0.055
	0.1	0.078	0.109	0.108	0.098	0.107	0.043	0.101	0.106	0.097	0.098

Source: author's own.

Table 2

Size estimates of conditional coverage *VaR* tests

Test	Significance Level	1% *VaR*					2.5% *VaR*				
		Sample size					Sample size				
		100	250	500	750	1000	100	250	500	750	1000
LR_{CC}	0.01	0.011	0.012	0.014	0.013	0.013	0.022	0.036	0.026	0.028	0.034
	0.05	0.018	0.025	0.025	0.031	0.026	0.036	0.068	0.086	0.120	0.133
	0.1	0.022	0.031	0.041	0.051	0.045	0.052	0.100	0.184	0.203	0.186
LB_5	0.01	0.031	0.046	0.074	0.076	0.066	0.041	0.021	0.019	0.014	0.015
	0.05	0.034	0.077	0.114	0.120	0.102	0.050	0.058	0.055	0.051	0.052
	0.1	0.056	0.079	0.118	0.164	0.145	0.113	0.094	0.090	0.087	0.094

Source: author's own.

appear suggest the superiority of the *Z* test over the standard Kupiec LR_{UC} approach. The *Z* rejection frequencies seem accurate including all significance levels and both *VaR* coverage levels. They are also rather stable over the sample sizes, though the choice of the low-level *VaR*, like the ones considered, should clearly go with samples larger than 100 observations. The LR_{UC} rejection frequencies, in turn, show that the LR_{UC} distribution tends to diverge markedly from the theoretical likelihood ratio distribution. This is especially visible for 1% *VaR* and small sample sizes – for 1% *VaR* the convergence of LR_{UC} rejection frequencies to the nominal significance levels seems to start only from 750 observations.

The differences between the compared conditional coverage tests – LB_5 and LR_{CC} – are more pronounced. The results indicate that LB_5 outperforms the conventional LR_{CC} procedure. Especially for 2.5% *VaR*, the rejection

frequencies obtained for this test are clearly closer to the nominal significance, and they show signs of convergence to the desired levels with increasing the sample size. As opposed to this, the LR_{CC} test tends to be oversized for large samples, with rejection frequencies exceeding the nominal test size more than twice. For 1% coverage the empirical rejection frequencies of both tests do not correspond to the assumed significance levels – the tests tend to under-reject (LR_{CC}) or over-reject (LB_5) correct risk models. Therefore, the results suggest that in order to ensure an accurate test level, it is advisable to perform the conditional coverage testing for 2.5% *VaR*. Moreover, to reduce the type I error it is advisable to replace the Markov-chain-based *LR* statistic with the Ljung-Box statistic.

The power study, aimed at evaluating the test's ability to detect incorrect risk models, involved violation of the iid Bernoulli assumption. Relevant simulations were conducted in two stages, where the experiments subsequently violated unconditional and conditional coverage property. The false unconditional coverage was implemented through generating random Bernoulli numbers with the parameter π_1 set to values different than the *VaR* tolerance of 1% or 2.5%. As this type of experiment is dedicated to checking the unconditional coverage property, it was called the uc experiment. In the second stage the underlying processes violated the conditional coverage property. This was implemented through two types of experiments, called the cc experiments. Both cc experiments were aimed at generating serially dependent *VaR* failures, however were done in two ways. In the first type of the cc experiment, the focus was on controlling the scale of violation of the conditional coverage property, hence this experiment type is referred to as theoretically-oriented. The second type of the cc experiment strived to be as close as possible to the real market conditions, so this experiment type is viewed as practically-oriented.

In the first cc experiment type, the focus is on the scale of violating the conditional coverage property, which, in this case, is the same as the distance from the null. The author wanted to control this distance and treat it as the experiment parameter, and then observe how the test power changes when manipulating this parameter. In such experiments one needs a way to measure how much the conditional coverage property is violated. To achieve this, the author resorted to the basic *GARCH*-normal model, where the scale of violating the conditional coverage property can be judged from the volatility clustering, which in turn, can be measured by the autocorrelation of the squared returns. The dependence of failures is achieved by using a constant *VaR* level, based on the unconditional distribution of the returns. In this variant of the experiment, the author chose the following specification of the *GARCH* model:

$$\begin{aligned} R_t &= \sqrt{h_t} Z_t, \quad Z_t \sim N(0,1) \\ h_t &= \omega + \alpha R_{t-1}^2 + \beta h_{t-1} \end{aligned} \tag{6}$$

In accordance with the idea behind this type of experiment, specification (6) allows for the analytical calculation of autocorrelations of the squared returns, enabling to study the power of the test as a function of a controlled parameter of a return distribution. Under (6), the first order autocorrelation of the squared returns ρ is given by

$$\rho = \frac{\alpha^2 \beta}{1 - 2\alpha\beta - \beta^2}, \tag{7}$$

and the autocorrelations decline exponentially, with the decay factor $\alpha + \beta$. However, if the fourth moment of Z_t is not finite, the autocorrelations are time-varying. To prevent this, the model needs to satisfy the condition $(\alpha + \beta)^2 + 2\alpha^2 < 1$. This is ensured by fixing parameters ω and β at levels 0.01 and 0.6, respectively, and setting ρ to 0.1, 0.3 and 0.5 in subsequent variants of the experiment. The α parameter is set to such a value that ensures the desired level of ρ. This one obtains the simulation experiment that enables to explicitly control the volatility clustering.

In the second type of the cc experiments, the study aimed at closely mimicking the real-life conditions. For this reason, the more complex *GARCH* specifications were chosen, which represent various possible features of the financial data. The focus was on checking the test performance in specific conditions like non-linearity, non-normality, getting close to non-stationarity ($\alpha + \beta$ close to one), lack of the volatility clustering or the presence of the serial correlation in the mean equation instead of the variance equation. In the choice of the specific models matching the real data, the author followed previous studies by Berkowitz et al. (2011) and Du (2016), and used the following *N-GARCH* specification for the models, numbered from 1 to 4:

$$\begin{aligned} R_t &= \sqrt{h_t} Z_t \sqrt{\frac{d-2}{d}}, \quad Z_t \sim t(d), \\ h_t &= \omega + \alpha h_{t-1} \left(\sqrt{\frac{d-2}{d}} Z_{t-1} - \theta \right)^2 + \beta h_{t-1}, \end{aligned} \tag{8}$$

with the parameters

$$\begin{aligned} &\omega = 0.5469, \quad \alpha = 0.1552, \quad \beta = 0.7495, \\ &\theta = \text{-}0.245, \quad d = 3.808, \end{aligned} \tag{9}$$

(model 1)

(model 2) $$\omega = 0.2154, \quad \alpha = 0.0524, \quad \beta = 0.9284, \quad \theta = 0.5031, \quad d = 3.3183, \tag{10}$$

(model 3) $$\omega = 0.2127, \quad \alpha = 0.0261, \quad \beta = 0.8728, \quad \theta = -0.9616, \quad d = 6.9117, \tag{11}$$

(model 4) $$\omega = 1.6532, \quad \alpha = 0.0723, \quad \beta = 0.9153, \quad \theta = 0.0928, \quad d = 4.7017, \tag{12}$$

and the AR(2) return specification for the fifth model in this set of experiments:

(model 5) $$R_t = aR_{t-1} + bR_{t-2} + \sigma Z_t, \quad Z_t \sim N(0,1), \quad a = 0.5, \quad b = 0.3, \quad \sigma^2 = 0.1. \tag{13}$$

Generating data from the above processes required dispensing with the explicit control of the scale of the volatility clustering. Thus, the power estimates from these experiments cannot be compared to each other and cannot be assessed in relation to the distance from the null. The interpretation of their results can only rely on the fact that any of the above representations, combined with the constant *VaR* corresponding to the unconditional return distribution, produces the clusters of *VaR* failures.

The power results show that the test ability to detect incorrect *VaR* models differs considerably with respect to the *VaR* coverage (Tables 3 to 5). Testing based on 2.5% *VaR* seems possible even for samples of 250 observations, while inference based on 1% *VaR* appears feasible only for large samples. Recommendable sample sizes, for such a low coverage level, start with 750 observations.

Table 3

Power estimates of unconditional coverage *VaR* tests

Test	π_1	1% *VaR*					π_1	2.5% *VaR*				
		Sample size						Sample size				
		100	250	500	750	1000		100	250	500	750	1000
LR_{UC}	0.1%	0.001	0.000	0.306	0.483	0.631	0.5%	0.002	0.362	0.904	0.976	0.994
	0.5%	0.034	0.014	0.207	0.250	0.433	1.5%	0.025	0.085	0.375	0.426	0.561
	2%	0.371	0.235	0.546	0.639	0.787	3.5%	0.290	0.179	0.308	0.388	0.452
	3%	0.615	0.626	0.929	0.979	0.997	4.5%	0.475	0.459	0.730	0.869	0.932
Z	0.1%	0.906	0.782	0.908	0.992	0.997	0.5%	0.602	0.652	0.985	0.999	1.000
	0.5%	0.627	0.283	0.289	0.485	0.440	1.5%	0.225	0.110	0.375	0.427	0.561
	2%	0.457	0.395	0.543	0.731	0.781	3.5%	0.168	0.276	0.308	0.388	0.452
	3%	0.633	0.764	0.929	0.979	0.997	4.5%	0.306	0.580	0.730	0.869	0.932

Source: author's own.

Table 4

Power estimates of conditional coverage *VaR* tests

Test	ρ	1% *VaR*					ρ	2.5% *VaR*				
		Sample size						Sample size				
		100	250	500	750	1000		100	250	500	750	1000
LR_{CC}	0.1	0.131	0.164	0.205	0.259	0.266	0.1	0.099	0.153	0.109	0.147	0.161
	0.3	0.242	0.400	0.543	0.638	0.714	0.3	0.278	0.395	0.462	0.607	0.708
	0.5	0.499	0.548	0.675	0.774	0.863	0.5	0.381	0.411	0.635	0.775	0.874
LB_5	0.1	0.120	0.189	0.334	0.374	0.407	0.1	0.125	0.229	0.319	0.384	0.438
	0.3	0.337	0.455	0.718	0.820	0.889	0.3	0.302	0.571	0.800	0.898	0.948
	0.5	0.467	0.601	0.813	0.920	0.965	0.5	0.446	0.696	0.905	0.968	0.987

Source: author's own.

Table 5

Power estimates of conditional coverage *VaR* tests

Test	Model	1% *VaR*					Model	2.5% *VaR*				
		Sample size						Sample size				
		100	250	500	750	1000		100	250	500	750	1000
LR_{CC}	1	0.181	0.120	0.099	0.092	0.116	1	0.082	0.075	0.097	0.110	0.098
	2	0.308	0.219	0.144	0.177	0.216	2	0.100	0.110	0.192	0.247	0.287
	3	0.354	0.271	0.212	0.260	0.319	3	0.123	0.152	0.263	0.353	0.404
	4	0.309	0.285	0.348	0.405	0.503	4	0.192	0.276	0.433	0.548	0.619
	5	0.364	0.415	0.548	0.662	0.760	5	0.331	0.559	0.756	0.868	0.938
LB_5	1	0.091	0.088	0.110	0.130	0.148	1	0.071	0.103	0.140	0.165	0.183
	2	0.107	0.148	0.211	0.302	0.375	2	0.084	0.190	0.343	0.479	0.568
	3	0.154	0.208	0.317	0.418	0.513	3	0.116	0.259	0.458	0.620	0.711
	4	0.264	0.348	0.463	0.596	0.693	4	0.236	0.421	0.643	0.776	0.857
	5	0.356	0.471	0.608	0.741	0.831	5	0.404	0.629	0.856	0.944	0.977

Source: author's own.

As in the size study, larger differences in the test quality are connected with testing the conditional coverage property rather than the unconditional coverage property. The relative assessment of the unconditional coverage tests – Z and LR_{UC} (Table 3) – indicates that these tests are comparable in terms of their power. Any observed differences, however, indicate the Z normal statistic as the more powerful than the standard Kupiec LR_{UC} approach.

The results from testing the conditional coverage by LB_5 and LR_{CC} show remarkable differences in the test quality. The LB_5 test clearly outperforms the LR_{CC} procedure in all cc experiments of type one (*GARCH*-normal-based,

with the volatility clustering controlled by ρ), with rejection frequencies often doubling those of LR_{CC} (Table 4). In these experiments, the LB_5 supremacy is most visible at short distance from the null. For example, in the 0.1 correlation experiment, the LB_5 rejection frequencies tend to double or even triple (depending on the *VaR* level) those of LR_{CC}. Further from the null, the LB_5 outperformance is most marked for small samples.

The above conclusions from the first set of the cc experiments were confirmed by the experiments of the second type (*AR* and *N-GARCH-Student-t*-based). These experiments were designed with the aim of closely reflecting the real time series on the proviso of not having any parameter to control the volatility clustering. Therefore, the power results cannot be compared among the models, and there is no a priori knowledge of what power to expect for the specific data generating processes. What can be compared, however, is the rejection frequencies obtained for the standard LR_{CC} and the proposed LB_5 (Table 5). In the vast majority of cases, the rejection frequencies of LB_5 exceed those of LR_{CC}. This regularity can be observed without any exception for the 2.5% *VaR* level, which includes both the *ARMA* model and all *N-GARCH Student-t* data generating processes. For 1% *VaR* the only exceptions occur for the shortest series of 100 or 250 observations. The irregularities for 1% *VaR* and the shortest series can be explained by the small number of the observed *VaR* failures. For example, in the case of the 1% *VaR* and 100 observations, the expected number of *VaR* failures is 1. Such a low number of observations hinders any statistical inference. Thus for 1% *VaR*, finding patterns connected with statistical methods requires longer series[3]. Starting from 500 observations, as before, the LB_5 test systematically outperforms the LR_{CC} procedure. In summing up the results from both experiment variants, the LB_5 test appears more effective than the standard approach.

4. BACKTESTING EMPIRICAL *VaR* FORECASTS

The empirical study, based on FTSE100, NIKKEI225 and S&P500 data, illustrates how the statistical properties of the examined tests translate into practical conclusions from the risk analysis. To this end, twelve leading time series models, used to forecast daily *VaR*, were evaluated subsequently by all the tests. The range of the models covered both parametric and non-parametric methods. Within the parametric framework, the basic constant variance models

[3] Due to the rarity of observed *VaR* failures when testing a low-level *VaR*, similar studies often use samples of at least 500 observations – see e.g. Angelidis et al. 2004, Escanciano and Olmo 2011, Totić et al. 2011, Pajhede 2017, Patton et al. 2019.

were followed by conditional variance models with various error term specifications. The study employed the normal distribution, the Student-*t* distribution as well as the Picks over Thresholds (*POT*) method (McNeil and Frey 2000), which, through the Extreme Value Theory, uses the Generalized Pareto Distribution (Balkema and de Haan 1974, Pickands 1975). The conditional variance was modelled through the *GARCH*-class processes (Bollerslev 1986). This ensures representation of the volatility clustering phenomenon. To allow also for an asymmetric volatility response, relating to upward and downward market trends, the asymmetric *GJR-GARCH* models were used (Glosten et al. 1993). Within nonparametric methods the author employed the historical simulation model and the filtered historical simulation technique (Barone-Adesi et al. 1998), with filtering based on *GARCH* or *GJR-GARCH* model residuals.

The choice of the above models matched the aims of this study in the sense of evaluating a range of models, characterized by various levels of complexity and flexibility. As the study did not focus on finding the best fit to the time series, but on assessing the *VaR* tests, the author mainly needed the models that differ in their predictive ability. These models were applied as tools to generate a series of *VaR* forecasts, subsequently used to conduct the tests. The results regarding the quality of the models were treated in the study as complementary, while the key conclusions were based on the consistency among the tests, or the differences they showed in evaluating the *VaR* forecasts. For these reasons, the range of the time series model started with the most naive ones (like homoscedastic or the historical simulation models) and ended with the specifications regarded as flexible and showing high predictive ability (like the *GJR-GARCH*[4] models with the *t*-distribution or the *GJR-GARCH* models combined with the distribution based on the Extreme Value Theory).

The FTSE100, NIKKEI225 and S&P500 data were chosen to represent financial returns because these indexes are commonly used in similar research, showing typical features of the financial market. Such choice allowed to use the results of previous research and to compare the conclusions. One of the extensive studies, including these indexes, was carried by Angelidis et al. (2004). Based on the period 1987-2002, with 484 models for each index and two *VaR* levels, they showed that the *GARCH* models are unquestionable leaders in predicting *VaR*, but the right model choice strongly depends on the market specificity. The only attainable general conclusion, not depending on

[4] The predictive ability of various time series models strongly depends on the underlying data, however the good performance of the *GJR-GARCH* model in comparison to other specifications, like the basic *GARCH, fGARCH-TGARCH or EGARCH* was shown in several studies (e.g. for the S&P data in Hung-Chun and Jui-Cheng 2010 and Bilyk et al. 2020).

the particular market, was that the asymmetric models perform better than the others. This study was based on the standard Kupiec and Christoffersen procedure, so it was extended by including more powerful tests. Another similar study, based on the FTSE100 data from the 1997-2011 period, recommended the use of the *GARCH-POT* models that originated from the Extreme Value Theory (Totić et al. 2011). This study focused, however on testing only the unconditional coverage property. On the other hand, a recent study, which included the FTSE100, NIKKEI225 and S&P500 data from 1990 till 2016, concentrated on testing the conditional coverage property (Patton et al. 2019). Although this study placed greater focus on *ES* as a measure of risk than *VaR*, it showed the better predictive ability of the nonparametric *GARCH* models (represented in this study by the *FHS* method) than that of the parametric ones. This study, however, did not consider the combination of the *GARCH* models with the *POT* method, as is done here, but placed emphasis on the GAS (Generalized Autoregressive Score) approach. What is important, it referred to the early Christoffersen's *VaR* test, supporting the need to replace it with other testing methods.

As in majority of similar studies, the author's empirical analysis was based on the daily close-to-close log returns. In order to mimic the real-life decision-making process, where a standard sample of daily data covers a yearly period or its multiple, it was decided to perform the study on 4-year samples. As a result, there were around 1000 observations in each sample, which corresponds to the largest sample size examined in the simulation study. On the one hand this matches the risk management practice, and on the other, it provides a relatively wide range of data in one sample. Such a sample length also goes in line with other similar studies. For example, Angelidis et al. (2004), who put great emphasis on the sample choice, showing that best *VaR* predictions (for 1% *VaR* and *GARCH* models with normal or Student-*t* innovations) are attainable from samples of 1000 observations. Another element of the business practice is to repeat the testing with a fixed frequency, such as weekly, monthly or yearly. The practical choice of the frequency depends on the potential impact of the risk exposure on the company operations. As an effect of the sample choice and the frequency choice, the real-life samples usually overlap and any changes in the market conditions can be observed from a series of subsequent samples. To follow this practice, this process was repeated for several samples of the same 4-year length, however moving the testing window in such a way that the neighbouring samples do not overlap. This allowed to cover a wider time range and obtain more diversified samples, deemed relevant for the purposes of checking the capacity of risk management tools. An important element of the sample choice was to rely on predefined

time intervals instead of using any statistical techniques of dividing the data into specific subperiods. This was important for two reasons. First, it served to assess the test's ability of finding incorrect risk forecasts based on the pre-established periods. Therefore some fixed data and a collection of risk models were needed, with a varied potential of matching these data. If a change in the volatility regime occurs, it is expected that the tests find the unsuitability of the applied risk models. Second, the aim was to follow the real-life procedures where the risk model testing is carried out periodically, usually based on calendar periods, without any a priori knowledge of shifts between the market regimes. Thus the author decided to use the samples: 2008-2011, 2012-2015, 2016-2019. Such a subsample choice has the additional advantage of the oldest subsample going back as far as the subprime mortgage crisis of 2008, which gave the possibility to empirically evaluate the test performance in the extreme conditions experienced in the recent past. The standard choice of the intuitive periods corresponding to the calendar years resulted in limiting the time series to the end of 2019. However, in order to fully use the various market situations experienced recently, the study was extended to the middle of 2020, hence including one more sample of the same length as all the others. It partly overlaps with the 2016-2019 sample but differs from it substantially, as it covers the outbreak of the COVID-19 pandemic. The latter sample goes from the middle of 2016 to the middle of 2020. In this way the author obtained four samples, which seem highly diverse, as shown by the descriptive statistics (Table 6).

Table 6

Descriptive statistics of S&P500, FTSE100 and NIKKEI225 daily returns

Index	Period	Mean	Std. deviation	Minimum	Maximum	Skewness	Kurtosis
FTSE100	2008-2011	-0.0001	0.016	-0.093	0.094	-0.06	8.20
	2012-2015	0.0001	0.009	-0.048	0.035	-0.23	5.05
	2016-2019	0.0002	0.008	-0.035	0.035	-0.16	5.35
	Mid-2016 to mid-2020	-0.0001	0.010	-0.115	0.087	-1.61	26.67
NIKKEI225	2008-2011	-0.0006	0.0198	-0.1211	0.1323	-0.49	10.11
	2012-2015	0.0008	0.0136	-0.0760	0.0743	-0.35	5.87
	2016-2019	0.0002	0.0118	-0.0825	0.0691	-0.46	10.08
	Mid-2016 to mid-2020	0.0004	0.0118	-0.0627	0.0773	-0.05	9.92
S&P500	2008-2011	-0.0002	0.0181	-0.0947	0.1096	-0.22	8.75
	2012-2015	0.0005	0.0081	-0.0402	0.0383	-0.26	4.87
	2016-2019	0.0004	0.0080	-0.0418	0.0484	-0.64	7.75
	Mid-2016 to mid-2020	0.0004	0.0125	-0.1277	0.0897	-1.20	27.39

Source: author's own.

A common observation for all the indexes is that the first sample, 2008-2011, including the subprime mortgage crisis and its spillover effects, clearly stands out. It represents the crisis-driven behaviour, which manifests itself in high volatility and excess kurtosis. The market crash resulted also in low means, extremely low minimums and relatively high maximums. For FTSE100 and S&P500, such behaviour is also highly evident in the last sample, mid-2016 to mid-2020, including the outbreak of the COVID-19 pandemic. For these two indexes the volatility, skewness and kurtosis went down in the 2012-2015 and 2016-2019 samples, which therefore were initially regarded as representative for the usual market conditions. NIKKEI225 differs from the above observations in the sense that the COVID-19 sample does not stand out so clearly, and the preceding periods also show signs of the high volatility regime. This can be observed also from the NIKKEI225 time series plots, which show large volatility clusters not just in the neighbourhood of the subprime mortgage or the COVID-19 crises (Figure 1).

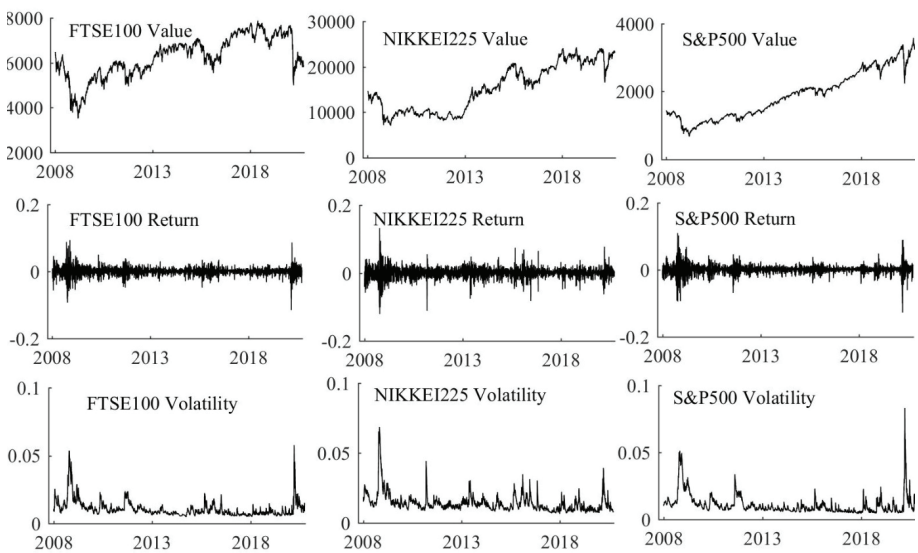

Fig. 1. S&P500, FTSE100 and NIKKEI225 values, returns and volatility

Source: author's own.

The backtesting exercise performed for the three indexes was aimed at illustrating the differences in the conclusions from the risk analysis, attributable to the test choice. The *p*-values obtained for all examined stock market indexes (Tables 7 to 9) show that these differences are minor when testing unconditional

coverage property through LR_{UC} and Z. On the other hand, testing conditional coverage by the means of LR_{CC} and LB_5 statistics reveals that the test choice markedly influences the outcomes. This conclusion is in line with the results of the simulation study.

Backtesting a risk model with respect to the unconditional coverage (LR_{UC} and Z tests) informs whether the overall number of *VaR* violations produced by the examined model, corresponds to the assumed *VaR* tolerance level. The general picture from backtesting risk models by Z and LR_{UC} is that both unconditional coverage tests provide comparable conclusions. There are

Table 7

p-values of *VaR* tests for FTSE100 index

VaR model	2008-2011				2012-2015			
	LR_{UC}	Z	LR_{CC}	LB_5	LR_{UC}	Z	LR_{CC}	LB_5
HS	0.607	0.301	0.001**	0.000**	0.497	0.253	0.010**	0.000**
GARCH-FHS	0.624	0.315	0.521	0.273	0.466	0.228	0.052*	0.023**
GJR-GARCH-FHS	0.624	0.315	0.521	0.273	0.466	0.228	0.052*	0.023**
Normal	0.000**	0.000**	0.017**	0.000**	0.267	0.126	0.076*	0.000**
GARCH-Normal	0.029**	0.010**	0.670	0.710	0.096*	0.040**	0.425	0.391
GJR-GARCH-Normal	0.007**	0.002**	0.660	0.404	0.195	0.089*	0.799	0.666
Student-t	0.000**	0.000**	0.039**	0.000**	0.195	0.089*	0.090*	0.000**
GARCH- Student-t	0.144	0.063*	0.789	0.416	0.195	0.089*	0.350	0.278
GJR-GARCH- Student-t	0.004**	0.001**	0.694	0.625	0.267	0.126	0.799	0.655
POT	0.001**	0.000**	0.002**	0.000**	0.025**	0.019**	0.211	0.000**
GARCH-POT	0.485	0.247	0.478	0.201	0.952	0.476	0.143	0.151
GJR-GARCH-POT	0.936	0.468	0.252	0.348	0.638	0.322	0.523	0.269
	2016-2019				Mid-2016 to mid-2020			
	LR_{UC}	Z	LR_{CC}	LB_5	LR_{UC}	Z	LR_{CC}	LB_5
HS	0.735	0.369	0.119	0.116	0.082*	0.033**	0.005**	0.000**
GARCH-FHS	0.892	0.446	0.139	0.143	0.414	0.201	0.060*	0.019**
GJR-GARCH-FHS	0.892	0.446	0.139	0.143	0.414	0.201	0.060*	0.019**
Normal	0.302	0.144	0.067*	0.010**	0.005**	0.001**	0.023**	0.000**
GARCH-Normal	0.113	0.048**	0.398	0.684	0.055*	0.021**	0.159	0.327
GJR-GARCH-Normal	0.302	0.144	0.794	0.581	0.023**	0.007**	0.572	0.790
Student-t	0.113	0.048**	0.111	0.010**	0.001**	0.000**	0.051*	0.000**
GARCH- Student-t	0.515	0.253	0.743	0.649	0.119	0.051*	0.413	0.572
GJR-GARCH- Student-t	0.223	0.103	0.325	0.394	0.001**	0.000**	0.043**	0.103
POT	0.238	0.128	0.051*	0.000**	0.233	0.108	0.002**	0.000**
GARCH-POT	0.452	0.231	0.453	0.778	0.812	0.405	0.185	0.253
GJR-GARCH-POT	0.892	0.446	0.264	0.703	0.531	0.262	0.754	0.608

Source: author's own.

Table 8

p-values of *VaR* tests for NIKKEI225 index

VaR model	2008-2011				2012-2015			
	LR_{UC}	Z	LR_{CC}	LB_5	LR_{UC}	Z	LR_{CC}	LB_5
HS	0.365	0.176	0.008**	0.000**	0.371	0.179	0.057*	0.000**
GARCH-FHS	0.610	0.308	0.475	0.223	0.277	0.130	0.296	0.646
GJR-GARCH-FHS	0.610	0.308	0.475	0.223	0.277	0.130	0.296	0.646
Normal	0.027**	0.009**	0.042**	0.000**	0.043**	0.016**	0.157	0.000**
GARCH-Normal	0.027**	0.009**	0.510	0.002**	0.201	0.092*	0.083*	0.174
GJR-GARCH-Normal	0.010**	0.003**	0.592	0.560	0.201	0.092*	0.331	0.497
Student-t	0.001**	0.000**	0.139	0.000**	0.066**	0.026**	0.135	0.000**
GARCH- Student-t	0.096*	0.040**	0.389	0.089*	0.201	0.092*	0.083*	0.174
GJR-GARCH- Student-t	0.027**	0.009**	0.700	0.507	0.098*	0.041**	0.116	0.194
POT	0.027**	0.009**	0.001**	0.000**	0.615	0.304	0.038**	0.000**
GARCH-POT	0.753	0.375	0.165	0.005**	0.201	0.092**	0.083*	0.174
GJR-GARCH-POT	0.911	0.455	0.608	0.056*	0.201	0.092**	0.331	0.497
	2016-2019				Mid-2016 to mid-2020			
	LR_{UC}	Z	LR_{CC}	LB_5	LR_{UC}	Z	LR_{CC}	LB_5
HS	0.792	0.397	0.113	0.000**	0.286	0.135	0.000**	0.000**
GARCH-FHS	0.593	0.293	0.714	0.876	0.492	0.241	0.759	0.999
GJR-GARCH-FHS	0.593	0.293	0.714	0.876	0.492	0.241	0.759	0.999
Normal	0.358	0.172	0.778	0.001**	0.071*	0.028**	0.001**	0.000**
GARCH-Normal	0.138	0.061*	0.798	0.847	0.150	0.066**	0.382	0.186
GJR-GARCH-Normal	0.593	0.293	0.714	0.779	0.622	0.308	0.727	0.867
Student-t	0.000**	0.000**	0.434	0.001**	0.000**	0.000**	0.004**	0.000**
GARCH- Student-t	0.138	0.061*	0.798	0.847	0.150	0.066*	0.382	0.186
GJR-GARCH- Student-t	0.096*	0.040**	0.781	0.856	0.150	0.066*	0.790	0.715
POT	0.185	0.102	0.336	0.363	0.172	0.096*	0.046**	0.000**
GARCH-POT	0.593	0.293	0.714	0.778	0.380	0.183	0.277	0.360
GJR-GARCH-POT	0.888	0.444	0.633	0.725	0.920	0.460	0.606	0.781

Source: author's own.

however a few cases when Z rejects the models that LR_{UC} allows, which suggests that the normal Z statistics proved more powerful at detecting incorrect models. Such cases happened for all three indexes, most often however for NIKKE225. A vivid example is the 2012-2015 NIKKEI225 sample, where Z rejects nearly all parametric specifications, whereas LR_{UC} admits most of them. In general, however, both tests point out similar models as acceptable for predicting risk for all three indexes. In particular, both tests classify the majority of parametric models, apart from those belonging to the *POT* class, as incorrect in the high volatility regime. This is especially evident

Table 9

p-values of *VaR* tests for S&P500 index

VaR model	2008-2011				2012-2015			
	LR_{UC}	Z	LR_{CC}	LB_5	LR_{UC}	Z	LR_{CC}	LB_5
HS	0.003**	0.001**	0.100	0.000**	0.980	0.490	0.145	0.000**
GARCH-FHS	0.381	0.197	0.333	0.000**	0.708	0.352	0.194	0.216
GJR-GARCH-FHS	0.381	0.197	0.333	0.000**	0.708	0.352	0.194	0.216
Normal	0.000**	0.000**	0.241	0.000**	0.252	0.118	0.318	0.000**
GARCH-Normal	0.000**	0.000**	0.296	0.026**	0.129	0.056*	0.391	0.157
GJR-GARCH-Normal	0.000**	0.000**	0.224	0.339	0.129	0.056*	0.785	0.796
Student-t	0.000**	0.000**	0.002**	0.000**	0.000**	0.000**	0.431	0.019**
GARCH- Student-t	0.000**	0.000**	0.382	0.042**	0.339	0.162	0.284	0.087*
GJR-GARCH- Student-t	0.000**	0.000**	0.247	0.418	0.129	0.056*	0.785	0.796
POT	0.000**	0.000**	0.241	0.000**	0.001**	0.002**	0.004**	0.000**
GARCH-POT	0.457	0.223	0.182	0.006**	0.708	0.352	0.194	0.216
GJR-GARCH-POT	0.457	0.223	0.182	0.615	0.980	0.490	0.614	0.717
	2016-2019				Mid-2016 to mid-2020			
	LR_{UC}	Z	LR_{CC}	LB_5	LR_{UC}	Z	LR_{CC}	LB_5
HS	0.512	0.251	0.243	0.000**	0.009**	0.002**	0.080*	0.000**
GARCH-FHS	0.944	0.472	0.646	0.000**	0.528	0.260	0.755	0.000**
GJR-GARCH-FHS	0.944	0.472	0.646	0.000**	0.528	0.260	0.755	0.000**
Normal	0.021**	0.007**	0.057*	0.000**	0.000**	0.000**	0.030**	0.000**
GARCH-Normal	0.789	0.394	0.187	0.017**	0.231	0.107	0.339	0.006**
GJR-GARCH-Normal	0.789	0.394	0.686	0.940	0.167	0.075*	0.376	0.574
Student-t	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
GARCH- Student-t	0.789	0.394	0.187	0.017**	0.312	0.149	0.305	0.003**
GJR-GARCH- Student-t	0.159	0.071*	0.379	0.351	0.023**	0.007**	0.208	0.278
POT	0.300	0.142	0.000**	0.000**	0.003**	0.001**	0.000**	0.000**
GARCH-POT	0.896	0.448	0.603	0.012**	0.528	0.260	0.755	0.001**
GJR-GARCH-POT	0.590	0.298	0.517	0.785	0.808	0.403	0.684	0.604

Source: author's own.

in 2008-2011 subprime mortgage crisis samples and, to a lesser extent, in the mid-2016 and mid-2020 COVID-19 samples. The admitted models in these highly volatile periods are based either on the nonparametric historical simulation method (HS or FHS class) or the POT method originating from the Extreme Value Theory. This shows that, when turbulences occur, the popular distributional assumptions of normality or Student-t innovations tend to produce an excessive number of VaR violations. In calmer samples, the correct overall failure rate is attainable by most of the methods. Yet, judging by the p-values, the FHS or POT models seem to perform best in terms of the unconditional coverage.

The conditional coverage tests (LR_{CC} and LB_5) complement the overall *VaR* failure rate check by enquiring into the dependence of failures in time. The procedures dedicated to this property are often deemed critical to financial stability as they potentially prevent catastrophic losses from occurring in series. In light of the results of the conditional coverage tests, the choice among the constant-variance, conditional-variance or asymmetric conditional-variance specifications turns out to be more important for forecasting risk than the distributional choice. First of all, the use of the *GARCH*-class models is indicated as crucial for preventing *VaR* failure dependence in time. Second, in some cases, the asymmetric *GJR-GARCH* models are strongly preferred. However, most importantly in view of the study's goals, testing the conditional coverage property reveals substantial differences in the conclusions, attributable to the chosen testing method.

For the FTSE100 index (Table 7) the LB_5 test generally classifies *GARCH*-class models as admissible. In the case of this index, the LB_5 test does not distinguish between the standard *GARCH* and the *GJR-GARCH* models with volatility asymmetry. This shows that the potential differences in the market behaviour relating to upward and downward trends, do not impact on the FTSE100 risk forecasts. This result is stable across the samples, which indicates that although the model parameters may change, the volatility regime does not affect the general patterns of investors' behaviour. Another conclusion is the preference towards the parametric models over the *FHS* ones, which is visible in two out of four samples. Compared to these outcomes, based on the LB_5 procedure, the standard Markov LR_{CC} test results seem more vague. In some cases the Markov test even fails to reject the most naive, constant-variance models (the *POT* model in 2012-2015 and the Student-*t* model in 2016-2019).

The LB_5 *p*-values from testing the NIKKEI225 (Table 8) conditional coverage show a different pattern of market behaviour in comparison to the FTSE100 index. Contrary to FTSE100, the risk model choice for this index seems to be driven by volatility regimes. The LB_5 test shows that in the 2008-2011 market crash sample, only the narrow class of the *GJR-GARCH* models is capable of producing accurate risk forecasts. Thus only these models have the potential to prevent occurring large losses clustered in time during extremely volatile periods. In other samples which do not include such extraordinary price movements, the more general class of the *GARCH* models turns out to be sufficient for predicting risk. Since the volatility asymmetry component of the *GJR* models appears crucial only for times of crisis, it appears that the investors' behaviour is influenced by the volatility regime. The high crisis-driven volatility appears to stimulate more violent reactions to

falling prices. This asymmetric volatility regime-specific effect is strong enough to affect the suitability of risk forecasting methods.

Similarly to the case of the FTSE100 index, backtesting the NIKKEI225 risk forecasts through the standard Markov LR_{CC} procedure provides a different picture than backtesting through LB_5. In most cases the LR_{CC} test is unable to point out any specific class of models. For the 2008-2011market crash sample it failed to reject the basic *GARCH*-class models, classified as incorrect by the LB_5 statistic. For the two subsequent samples it admits most of the models, failing to specify any approach recommendable for predicting risk. In particular, for 2016-2019 all the models are admitted.

The results from testing the conditional coverage for S&P500 by LB_5 are in line with those for NIKKEI225 (Table 9). In standard situations, as indicated by the 2012-2015 and 2016-2019 samples, the *GARCH* risk forecasts are sufficient to fulfil the requirement of the proper conditional coverage. However, to ensure that *VaR* violations are not serially correlated during the crises, the asymmetry volatility component needs to be taken into account. Thus the *GJR-GARCH*-class models are advisable for the 2008-2011 subprime mortgage crisis sample and the mid 2016 to mid-2020 COVID-19 sample. An additional observation for S&P500, which goes in line with the FTSE100 results, is the preference towards the parametric specifications over the historical simulation-based methods. This is clear from all samples apart from the calmest 2012-2015 one.

As previously, the S&P500 results are test-specific. The risk analysis based on the Markov LR_{CC} statistic gives different conclusions. Under all volatility regimes it admits models from various classes, failing to characterize the market specificity.

The outcomes of the conditional coverage tests for all the indexes, combined with the initial results from testing unconditional coverage property, indicate that the *GJR-GARCH-POT* model performs best overall in terms of forecasting daily risk for stock prices. It seems most flexible as it is classified as accurate in light of both properties, for all indexes and under all volatility regimes. This most general result stays in line with the results of other similar studies that assessed *VaR* predictability for periods including major stock crashes (e.g. Angelidis et al. 2004, Totić et al. 2011). Referring to the study's goals, an important fact is that this conclusion can be deduced only from the LB_5 results, in particular not attainable by the Markov LR_{CC} test.

With regard to the market specificity, the combined results from LB_5 test for the three examined indexes allow for distinguishing two distinct patterns of investors' behaviour connected with undergoing the financial crises. While in the London market, as judged by the FTSE100 index, ways of predicting

risk seem insensitive to the volatility regimes, the other markets appear to be strongly affected by the crises. According to the NIKKEI225 and S&P500 results, in the crises conditions the general class of *GARCH* models needs to be narrowed to the models with the volatility asymmetry. Otherwise, the violations of *VaR* tend to cluster in time, which may result in large losses occurring one by one. Since the volatility asymmetry is essential for periods with extraordinary price movements, the financial crisis in the New York and Tokyo markets seems to affect investors' behaviour in such a way that it stimulates violent reactions to downward price movements. A crucially important fact is that these conclusions about market specificity strongly depend on the backtesting method. The results demonstrate that, contrary to the LB_5 autocorrelation test, the Markov LR_{CC} test used commonly in the industry, fails to explain the individual nature of the markets.

SUMMARY AND CONCLUSIONS

The study dealt with the methods of evaluating risk forecasts. The author referred to the Basel framework, which recommends testing risk models based on the *VaR* measure, and inquired into the statistical properties of the *VaR* tests. In order to enhance their accuracy and efficiency, the study replaced standard risk-management-dedicated tests with other econometric methods applicable to the binary *VaR* failure process, and decomposed *VaR* model evaluation into testing the unconditional and conditional property. With respect to the verification of the unconditional coverage property, the study utilized the convergence of the binomial distribution to the normal one, while for the conditional coverage, it employed the Ljung-Box χ^2 statistic. Therefore, it was proposed to use the well-established econometric methods instead of the methods developed specifically for the purposes of risk management.

In accordance to the Basel rules, the author examined the test properties on two low significance levels, and the simulations confirmed the superiority of the proposed procedures over the industry standards, in terms of their power. The results of the simulations were used in the empirical study, which demonstrated the advantages of the proposed approach. The study was designed with a view to showing the differences in the risk analysis attributable to the choice of the backtesting method. To provide a relevant setting for evaluating risk management tools, the study used data on three leading stock market indexes, various volatility regimes and twelve mainstream risk models to generate *VaR* forecasts. The application of the proposed methods to the *VaR* failure series provided evidence that more powerful tests, in comparison to the

standard risk management procedures, give a more insightful view of the market behaviour.

Contrary to the standard approach, the proposed procedures allowed for distinguishing models that best suit risk management in various market conditions. This, in turn, enabled to define market-specific patterns of investors' behaviour connected with changing volatility conditions. Under usual conditions, the general class of *GARCH* models was pointed out as sufficient in terms of predicting risk. However, the inclusion of the financial crises into the sample implied the need for more specific methods. In the market crash periods only the narrow class of the nonparametric filtered historical simulation models or the *POT* models prevented risk underestimation. The requirement that the *VaR* failures should not group in time further narrowed the range of acceptable models only to those that combine the *POT* method with the *GARCH* volatility specification. Moreover, for the New York and Tokyo markets, judging by their leading indexes, the asymmetry volatility component was vital in the sense that it prevented clustering of extraordinary large losses. This indicated the *GJR-GARCH-POT* model as the most flexible, in the sense of being suitable for predicting risk in the widest variety of market conditions. A comparison of these results to the outcomes of the standard *VaR* testing procedure demonstrated that the proposed methods were more effective in detecting incorrect risk models. Two general implications follow from this comparison: first, the proposed tests better characterize the specificity of the market behaviour and second, more importantly, they have better potential to secure the stability of the institutions operating in the financial markets. Thus, the results motivated the author to recommend these methods for institutional risk management systems as a replacement for the usual LR-based procedures. Moreover, these conclusions may also be used as guidance by the supervisory bodies in creating recommendations for risk managers.

The conclusions, formulated in the most general form, suggest the superiority of the well-established econometric methods over the standard risk management tools. In more detail, however, the main improvements were achieved by replacing the Markov-chain Christoffersen's framework with the testing based on the LB statistic. The author treated this not only as an indication of the recommendable *VaR* testing approach, but also as guidance for further developments. Following the idea behind the *LB* statistic, one of the directions for future research may be to search for more advanced ways of using the autocorrelation function. Their potential may lie, among others, in employing the spectral theory which allows for utilizing the same information as included in autocorrelations, but modified by means of the Fourier transform. Such a transform, by using the same information in a different way, may

improve the power properties. The use of the spectral theory gives a wide range of possibilities connected with new testing statistics. Such an approach was proposed in the *VaR* testing context by Berkowitz et al. (2011), but has been studied so far only in a very limited scope, based on two chosen statistics. One more proposition of utilizing the autocorrelation function in testing *VaR* is the recent modification of the LB statistic by Miettinen et al. (2020). Unlike the basic LB test, this modification takes into account the presence of the volatility clustering. In this modification, the asymptotic variance of the test statistic is derived when assuming only the symmetry and finite fourth moments of the time series. When the time series has the volatility clustering, it introduces a multiplicative factor that helps to achieve the correct size of the test. Both this proposition and the one suggesting to utilize the spectral theory require extensive simulations and empirical verification, and thus are left for further research.

Another natural extension to this study is to verify the author's propositions with the use of multivariate *GARCH* processes. Such processes, exactly as with the univariate ones, allow to predict *VaR*. Indeed, one of the key practical advantages of *VaR* as a risk measure is the straightforward way it can be computed for portfolios of assets or portfolios of indexes. As a consequence, testing multivariate *VaR* models proceeds in an analogous way to testing univariate ones. Up to now, several studies have been conducted to test accuracy of the multivariate models like *VECH, BEKK, CCC-GARCH, DCC-GARCH* and asymmetric *DCC-GARCH* in the context of forecasting *VaR* (Morimoto and Kawasaki 2008, Caporin and McAleer 2014, Santos, Nogales and Ruiz 2013). These studies, however, were based on the standard risk management tests. Validating the multivariate models by means of the methods found relevant in the univariate case (and possibly other, improved tests based on autocorrelations) is an area viewed as an interesting subject for future studies.

REFERENCES

Angelidis, T., Benos, A., Degiannakis, S., *The use of GARCH models in VaR estimation*, "Journal of Statistical Methodology", Vol. 1, pp. 105–128, 2004.

Balkema, A., de Haan, L., *Residual life time at great age*, "The Annals of Probability", Vol. 2, pp. 792–804, 1974.

Barone-Adesi, G., Burgoin, F., Giannopoulos, K., *Don't look back*, "Risk", Vol. 11, pp. 100–104, 1998.

Basel Committee on Banking Supervision, *Minimum capital requirements for market risk*, http://www.bis.org/bcbs/publ/d352.pdf (accessed June 4, 2018), 2016.

Basel Committee on Banking Supervision, *High-level summary of Basel III Reforms,* https://www.bis.org/bcbs/publ/d424_hlsummary.pdf (accessed June 4, 2018), 2017.

Berkowitz, J., *Testing Density Forecasts with Applications to Risk Management*, „Journal of Business & Economic Statistics", Vol. 19, No. 4, pp. 465–474, 2001.

Berkowitz, J., Christoffersen, P., Pelletier, D., *Evaluating Value-at-Risk Models with Desk-Level Data*, "Management Science", Vol. 12, No. 57, pp. 2213–2227, 2011.

Bollerslev, T., *Generalized autoregressive conditional heteroscedasticity*, "Journal of Econometrics", Vol. 31, pp. 307–327, 1986.

Bilyk, O., Sakowski, P., Ślepaczuk, R., *Investing in VIX futures based on rolling GARCH models forecasts*, [in:] Working Papers No. 10/2020 (316), Faculty of Economic Sciences, University of Warsaw, Warsaw, 2020.

Candelon, B., Colletaz, G., Hurlin, C., Tokpavi, S., *Backtesting Value-at-Risk: a GMM duration-based test,* "Journal of Financial Econometrics", Vol. 9, No. 2, pp. 314–343, 2011.

Caporin, M., McAleer, M., *Robust Ranking of Multivariate GARCH Models by Problem Dimension*, "Computational Statistics and Data Analysis", Vol. 76, pp. 172-185,2014.

Christoffersen, P., *Evaluating Interval Forecasts*, "International Economic Review", Vol. 39, No. 4, pp. 841–862, 1998.

Christoffersen, P., Pelletier, D., *Backtesting Value-at-Risk: A Duration-Based Approach*, "Journal of Financial Econometrics", Vol. 1, No. 2, pp. 84–108, 2004.

Colletaz, G., Hurlin, C., Perignon, C., *The Risk Map: a New Tool for Risk Management,* "Journal of Banking & Finance", Vol. 37, No. 10, pp. 3843–3854, 2013.

Crnkovic, C., Drachman, J., *Quality Control in VaR,* [in:] Grayling, S. (ed.) *VaR: Understanding and Applying Value-at-Risk*, Risk Publications, London, 1997.

Du, Z., *Nonparametric bootstrap tests for independence of generalized errors*, "The Econometrics Journal", Vol. 19, pp. 55–83, 2016.

Engle, R.F., Manganelli, A., *CAViaR: Conditional Autoregressive Value-at-Risk by Regression Quantiles*, "Journal of Business & Economic Statistics", Vol. 22, pp. 367–381, 2004.

Escanciano, J. C., Olmo, J., *Robust Backtesting Tests for Value-at-risk Models*, "Journal of Financial Econometrics", Vol. 9, No. 1, pp. 132–161, 2011.

Glosten, L., Jagannathan, R., Runkle, D., *On the relation between the expected value and the volatility of the nominal excess return on stocks*, "Journal of Finance", Vol. 48, pp. 1779–1801, 1993.

Gordy, M.B., McNeil, A.J, *Spectral Backtests of Forecast Distributions with Application to Risk Management*, [in:] *Finance and Economics Discussion Series* 2018-021, Board of Governors of the Federal Reserve System, Washington, 2018.

Hung-Chun, L., Jui-Cheng, H., *Forecasting S&P100 stock index volatility: The role of volatility asymmetry and distributional assumption in GARCH models*, "Expert Systems with Applications", Vol. 37, No. 7, pp. 4928–4934, 2010.

Hurlin, Ch., Tokpavi, S., *Backtesting value-at-risk accuracy: a simple new test*, „Journal of Risk", Vol. 9, No. 2, pp. 19–37, 2007.

J. P. Morgan, *Riskmetrics*, Technical document, Morgan Guaranty Trust Company, 1994.

Kratz, M., Lok, Y.H., McNeil, A. J., *Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall*, "Journal of Banking & Finance", Vol. 88, pp. 393-407, 2018.

Kupiec, P., *Techniques for Verifying the Accuracy of Risk Measurement Models*, "Journal of Derivatives", Vol. 3, No. 2, pp. 174–184, 1995.

Ljung, G. M., Box, G.E.P., *On a Measure of Lack of Fit in Time Series Models*, "Biometrika", Vol. 65, pp. 297–303, 1978.

Morimoto, T., Kawasaki, Y., *Empirical Comparison of Multivariate GARCH Models for Estimation of Intraday Value at Risk*, Available at SSRN: http://dx.doi.org/10.2139/ssrn.1090807, 2008.

McNeil, J., Frey, F., *Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach*, "Journal of Empirical Finance", Vol. 7, 271–300, 2000.

Miettinen, M., Matilainen, M., Nordhausen, K., Taskinen, S., *Extracting Conditionally Heteroskedastic Components using Independent Component Analysis*, "Journal of Time Series Analysis", Vol. 41, pp. 293–311, 2020.

Pajhede, T., *Backtesting Value at Risk: A Generalized Markov Test*, "Journal of Forecasting", Vol. 36, No. 5, pp. 597–613, 2017.

Patton, A. J., Ziegel, J. F., Chen, R., *Dynamic semiparametric models for expected shortfall*, "Journal of Econometrics", Vol. 211, No. 2, pp. 388–413, 2019.

Pickands, J., *Statistical inference using extreme order statistics*, "The Annals of Statistics", Vol. 3, pp. 119–131, 1975.

Santos, A., Nogales, F., Ruiz, E., *Comparing Univariate and Multivariate Models to Forecast Portfolio Value-at-Risk*, "Journal of Financial Econometrics", Vol. 11, No. 2, pp. 400–441, 2013.

Totić, S., Bulajić, M., Vlastelica, T., *Empirical comparison of conventional methods and extreme value theory approach in value-at-risk assessment*, "African Journal of Business Management", Vol. 5, No. 33, pp. 12810–12818, 2011.

Received: December 2018, revised: October 2020